A Hierarchical Method for Scientific Knowledge Structure Discovery Postprint

Authors: Li Hui, Tian Yadan

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] This paper proposes a novel hierarchical method for scientific knowledge structure discovery, providing reference for optimizing the knowledge structure discovery process and improving knowledge organization forms. [Method/Process] The method utilizes the LDA topic model to construct a hierarchical scientific knowledge structure discovery framework, automatically determines the number of structure layers based on average inter-topic similarity, extracts literature subsets for each topic by automatically screening thresholds in the "document-topic" probability matrix, and finally employs a tree diagram to display the knowledge structure of the scientific domain, uncovering associations and inheritance relationships among knowledge, and compares it with the hierarchical topic model HLDA method. [Results/Conclusion] Through empirical research and comparative analysis, it is demonstrated that the proposed method yields superior knowledge structures with enhanced topic representativeness and improved computational efficiency, showing significant improvements over HLDA in terms of single-layer topic distinctiveness and inter-layer topic inheritance.

Full Text

Preamble

Vol. 62 No. 13, July 2018 *ChinaXiv Cooperative Journal*

A Hierarchical Discovery Method for Scientific Knowledge Structure

Li Hui, Tian Yadan School of Economics and Management, Xidian University, Xi'an 710126

Abstract

[**Purpose/Significance**] This paper proposes a novel hierarchical discovery method for scientific knowledge structure, providing a reference for optimizing the knowledge structure discovery process and improving knowledge organization forms. [**Method/Process**] The method utilizes the LDA topic model to construct a hierarchical scientific knowledge structure discovery approach. It automatically determines the number of hierarchical levels based on the average similarity between topics, automatically filters thresholds in the "document-topic" probability matrix to intercept literature subsets for each topic, and finally employs tree diagrams to display the knowledge structure of scientific fields, revealing the correlations and inheritances among knowledge points. The method is compared with the hierarchical topic model HLDA approach. [**Results/Conclusions**] Through empirical research and comparison, the proposed method demonstrates superior knowledge structure, stronger knowledge topic representation, and higher operational efficiency, showing significant improvements over the HLDA method in both single-layer topic differentiation and inter-layer topic inheritance.

Keywords: LDA; Cloud Computing; Hierarchical; Knowledge Structure
Classification Number: G254
DOI: 10.13266/j.issn.0252-3116.2018.13.012

The generalization, intersection, and penetration of scientific research have created a complex and interwoven landscape across various fields. The diversification of research content poses challenges for understanding and mastering the intrinsic structure of knowledge, creating an inevitable contradiction between vast knowledge and limited individual capacity. For scholars new to a particular field, comprehensively understanding its knowledge structure often requires substantial effort. Scattered knowledge points and unstructured information hinder the formation of knowledge structures and impede deep interdisciplinary integration. Therefore, constructing a scientific and rational knowledge structure is of great significance for research. However, current research on knowledge structure is mostly based on bibliometrics, employing multivariate statistical analysis or social network analysis methods that use keyword co-occurrence for simple knowledge extraction, focusing on discovering scientific hotspots while failing to fully reveal the intrinsic knowledge structure of scientific research.

To address the limitations of traditional research methods, this paper proposes a new hierarchical scientific knowledge structure discovery method. This approach uses feature words generated by the LDA topic model to map knowledge points, 挖掘其更深层次的语义信息，突破了传统方法仅根据共现词刻画知识热点的局限性，揭示了科学领域知识点间的多粒度层次关系，解决了知识结构发现中常常忽略知识间继承性的问题，更能体现知识组织的本质。研究者可以通过浏览层次化的知识结构全面、快速地了解领域概况和知识点分布情况，有利于减少阅读的盲目性。

2 Related Research

Early research on knowledge structure primarily approached from the perspective of "talent cultivation," discussing what knowledge structures teachers, librarians, or professionals in various fields should possess. After 2000, research began focusing on scientific literature, with increasingly broad application domains. Current methods for discovering knowledge structure can be broadly categorized into three types: multivariate statistical analysis-based methods, social network analysis-based methods, and topic model-based methods.

Multivariate Statistical Analysis (MSA) methods for knowledge structure discovery primarily classify research fields or predict development trends, successfully applied in supply chain, knowledge management, and other domains. MSA methods, derived from classical statistical theory, often use Multidimensional Scaling (MDS) to create two-dimensional knowledge maps, supplemented by Cluster Analysis (CA) or Factor Analysis (FA) results to determine the number and boundaries of knowledge points. MDS can map knowledge points to low-dimensional space through nonlinear transformation, CA can display relationships among knowledge points using dendrograms, while FA simplifies analysis by replacing all variables with fewer factors. Most studies comprehensively utilize these three methods to explore domain knowledge structures.

Social Network Analysis (SNA) methods for knowledge structure discovery initially relied on bibliometrics, using external document features such as authors, citations, and institutions to analyze cooperation and citation relationships in scientific fields, yielding abundant research results. As research progressed, studies using SNA methods to discover knowledge structures from internal document features like abstracts and keywords gradually emerged, applied in Western economic geography, library and information science, and other fields. Foreign scholars have also combined MSA and SNA methods to complement each other for knowledge structure discovery.

The commonality between these two approaches lies in constructing keyword co-occurrence matrices for clustering, visualizing knowledge points through two-dimensional maps or network graphs to identify core research groups and track domain development trajectories, with a focus on disciplinary frontiers and hotspots. However, MSA methods are insensitive to the overall properties of knowledge and connections between nodes, easily overlooking special nodes and small knowledge groups. SNA methods are susceptible to software limitations, may lose useful information during data conversion, have certain data scale restrictions, and involve significant human intervention. Both methods struggle to reflect deeper semantic relationships among knowledge points, a problem that topic models can effectively solve.

Topic model-based knowledge structure discovery methods primarily employ the Latent Dirichlet Allocation (LDA) model or its improved variants. These methods typically build corpora from scientific literature titles and abstracts, using topic models to extract latent knowledge topics and feature words representing

their content. In 2016, Wang Yuefen et al. used the LDA topic model to deeply explore the knowledge structure of the domestic knowledge flow field from a disciplinary classification perspective. In the same year, H.C. Chang utilized the LDA topic model to mine the knowledge structure of information security. Since knowledge topics obtained through LDA consist of series of feature words, this representation can reflect semantic information, and the distribution of each document across topics can be intuitively displayed through probabilities, effectively avoiding the singularity of MSA and SNA methods that only characterize knowledge topics through co-occurring word pairs, making it more suitable for large-scale knowledge topic extraction. However, most of these methods directly use the classic LDA model for topic extraction, resulting in only surface-level knowledge hotspots without a complete hierarchical representation of knowledge structure, ignoring hierarchical inheritance relationships among knowledge points. To address these issues, scholars have begun using the hierarchical topic model HLDA (Hierarchical Latent Dirichlet Allocation) for knowledge organization research, with applications in book internal topic organization and patent analysis. Nevertheless, this method still has problems such as significant topic overlap, sparse document distribution in leaf nodes, and difficulty controlling hierarchical structure and topic differentiation.

Therefore, to better conduct hierarchical scientific knowledge structure discovery and address the shortcomings of HLDA, this paper focuses on two key issues: how to determine the number of knowledge structure levels to ensure reasonable hierarchical granularity, and how to determine the scope of literature subsets to make extracted knowledge topics more accurate. When designing the model, we utilize similarity between knowledge topics to provide a reasonable basis for determining knowledge structure levels, and balance document scope and literature quality from the initial literature set that generates knowledge topics to maximize topic quality and optimize the knowledge structure discovery process. We compare our method with the classic HLDA method to evaluate their respective advantages and disadvantages, providing references for related research methods in knowledge structure discovery.

3 Hierarchical Knowledge Structure Construction

3.1 Research Framework

This paper conducts hierarchical knowledge structure discovery based on the LDA topic model. The research framework is shown in Figure 1 [Figure 1: see original paper], divided into three parts from top to bottom: data layer, logic layer, and presentation layer. The data layer involves literature collection and preprocessing, including merging and organizing collected literature sets, as well as general procedures such as word segmentation and stop word removal. The logic layer implements the hierarchical scientific knowledge structure discovery process, using the LDA model to mine preprocessed corpora, calculating average similarity between topics to help determine knowledge structure levels, and automatically filtering thresholds to determine literature subset scopes

for lower-level topics. The presentation layer maps potential knowledge topics based on feature words obtained from the logic layer, draws scientific knowledge structure tree diagrams, and facilitates scholars' quick understanding of domain knowledge and its distribution.

3.2 Literature Collection and Preprocessing

Literature collection first requires determining search queries, then retrieving documents from professional databases and exporting required records, and finally merging and summarizing literature data while removing records with incomplete bibliographic information (including titles, abstracts, etc.). For topic extraction, the quality of raw data is crucial. Collected literature data is unstructured and must be uniformly preprocessed before topic extraction, including word segmentation, stop word removal, and lemmatization to generate modeling files required for hierarchical scientific knowledge structure construction. The first line of the modeling file is the total number of documents in the corpus, with each document occupying one line. Since this paper uses both Chinese and English literature sets, preprocessing operations differ slightly: English does not require word segmentation, while Chinese does not require lemmatization.

3.3 Hierarchical Scientific Knowledge Structure Discovery Method

The logic layer is the core component of the hierarchical scientific knowledge structure discovery method, with the following specific implementation process:

Step 1: Input the preprocessed modeling file into the LDA model, set LDA model parameters (see Section 4.2.1) and the minimum average similarity between topics.

Step 2: Set the number of topics for layer i, run the LDA program to output "document-topic" probability distribution and "topic-word" probability distribution files, thereby obtaining layer i knowledge topics.

Step 3: Calculate the average similarity between topics and compare it with the previously set minimum average similarity value to determine the number of knowledge structure levels. If the result is smaller, execute Step 4; otherwise, proceed to Step 6. The algorithm description is provided in Section 3.3.1.

Step 4: Continue the hierarchical process by re-intercepting literature subsets for lower-level topics. The algorithm description is provided in Section 3.3.2.

Step 5: Let i = i + 1 and return to Step 2 to generate lower-level knowledge topics.

Step 6: At this point, the hierarchical termination condition is satisfied. To ensure good differentiation between topics, lower-level topics are no longer mined. All potential knowledge topics are then mapped to generate the hierarchical scientific knowledge structure diagram.

3.3.1 Determining Knowledge Structure Levels As the hierarchical knowledge structure deepens, knowledge granularity gradually becomes finer, and knowledge points under the same topic become increasingly similar. At this stage, continuing to extract topics at the next level becomes meaningless. Therefore, this paper designs an algorithm to determine knowledge structure levels. Here, a minimum average similarity value S_0 between topics needs to be set in advance (empirically set between 0 and 1).

The algorithm is based on the mature vector space model, which essentially maps feature words under topics into vector space to obtain feature vectors for each topic. The specific implementation is as follows: For layer i topics T and T $(1 \leq k \leq j, 1 \leq r \leq j, k \neq r)$, where k and r are topic numbers and j represents the total number of topics, the top 25 feature words under each topic are merged without duplication into a set (feature words under each topic are sorted in descending order by distribution probability, and the top 25 feature words can adequately represent the topic; therefore, this paper sets 25 feature words for each topic). This yields an n-dimensional vector, where n is the total number of feature words. Each topic is represented by a vector $V[v_1, v_2, ..., v]$, where V is the distribution frequency of the topic for feature word a. If the topic does not contain feature word a, then $V = 0$. This gives us the feature vector V for the topic.

The cosine similarity between two topic feature vectors is calculated to obtain inter-topic similarity. Finally, all topics at this layer are traversed to calculate the mean inter-topic similarity and compare it with the set value S_0 to determine the termination layer. The pseudocode for the entire algorithm is shown in Figure 2 [Figure 2: see original paper].

3.3.2 Determining Literature Subset Scope In topic models, each document belongs to various topics with certain probabilities. To refine topics and depict their lower-level knowledge structures, topics must be extracted again from documents belonging to the topic. The quality of the literature subset at this stage significantly impacts the extraction effectiveness of lower-level topics. Traditional methods generally set a numerical value or percentage based on experience to intercept literature sets, which is highly subjective. If documents with high probabilities are selected, lower-level topics have good inheritance from upper-level topics but may lose important topics generated by low-probability documents. If the document scope is expanded, the quality of the literature set decreases, program efficiency reduces, and upper and lower-level knowledge topics become very similar, failing to achieve the purpose of refinement.

Therefore, to find a balance between document scope and topic quality, by observing the "document-topic" probability matrix, we find that if a threshold λ can be found to satisfy two conditions: (1) ensuring all documents can be assigned to corresponding topics without loss due to too small an intercept threshold, with a small duplicate assignment rate; (2) documents assigned to each topic have high belonging probabilities to that topic, without affecting

topic quality due to too large an intercept threshold. Then such a threshold is the balance point we seek.

Taking Figure 3 [Figure 3: see original paper] as an example, in a 6$×$5 "document-topic" probability distribution matrix, we first filter out the maximum value of each row: 0.275, 0.3889, 0.9198, 0.6868, 0.1328, and 0.4841. Then we filter out the minimum value among these six values, 0.1328, using this value as the intercept threshold λ (marked in red). This ensures that intercepted documents have relatively high distribution probabilities under corresponding topics, guaranteeing topic quality and satisfying condition (2). After automatically filtering threshold λ, we intercept documents with probabilities greater than or equal to λ for each topic as the lower-level literature subset for that topic. The interception results (marked in green and red) are: under Topic0, documents 3 and 4; under Topic1, documents 1, 2, and 6; under Topic2, document 4; under Topic3, document 5; under Topic4, document 3. This assigns 6 documents to various topics with a low duplicate assignment rate without losing any documents, satisfying condition (1).

For a "document-topic" probability distribution matrix, we first filter out the maximum value of each row, then filter out the minimum value among these values, which can serve as the intercept threshold for lower-level topic literature subsets. The formula is: $\lambda =$ (w P(D , T)) (where represents the "take minimum" operation and represents the "take maximum" operation). The pseudocode for this algorithm is shown in Figure 4 [Figure 4: see original paper].

3.4 Generating Knowledge Structure Hierarchy Tree

The presentation layer primarily presents the final hierarchical scientific knowledge structure to users through a graphical interface. It maps potential knowledge topics using feature words under topics, and experts familiar with the field summarize appropriate topic representative words to help scholars better understand topic semantic information. Based on the summarized topic words, the scientific knowledge structure is drawn layer by layer to achieve a complete hierarchical representation, generating the knowledge structure hierarchy tree. This hierarchical structure clearly displays parallel and inheritance relationships among knowledge points, as well as coarse-grained and fine-grained divisions, facilitating scholars' comprehensive understanding of scientific domain knowledge.

The term "Cloud Computing" has received widespread attention from academia and industry since its proposal and remains a research hotspot both domestically and internationally in recent years. To comprehensively explore the development status of this field, help researchers understand its knowledge structure, and validate the proposed hierarchical scientific knowledge structure discovery method (denoted as HSKSD, Hierarchical Scientific Knowledge Structure Discovering Method), we selected Chinese and English literature in the cloud computing field as experimental data sources to mine knowledge topics, draw hierarchical

knowledge structures, and finally analyze experimental results compared with the classic HLDA method.

4 Experiments and Analysis

4.1 Data Overview and Preprocessing

4.1.1 Data Overview This paper conducted searches as shown in Table 1 , filtering out a small number of documents with incomplete bibliographic information, ultimately obtaining 6,115 Chinese literature records and 4,843 English literature records.

As shown in Figure 5 [Figure 5: see original paper], no documents were retrieved in Chinese or English for 2005 and 2006, indicating that cloud computing development officially originated in 2007. By 2016, after nearly a decade of development, the publication trend maintained steady growth, with Chinese publications exceeding 1,000 in 2013 and English publications exceeding 1,000 in 2015, demonstrating that the influence of the cloud computing field is gradually expanding, with foreign research showing a trend of surpassing Chinese research after 2016.

4.1.2 Data Preprocessing The preprocessing stage processes raw literature corpora to generate data files required for modeling. Different preprocessing methods are adopted for different literature set types, with specific operations shown in Table 2 .

4.2 Experimental Settings and Results Analysis

4.2.1 Experimental Settings Since LDA-related algorithms are relatively mature, detailed explanations are omitted here. Parameter settings refer to literature [21] and empirical values from experimental summaries. Parameters α and β control the distribution of topics and words, with specific explanations provided in Table 3 .

HLDA modeling steps and parameter settings refer to literature [22]. To facilitate subsequent evaluation and make the knowledge structure levels and topic numbers generated by both methods as close as possible, after multiple experiments, specific parameter settings are shown in Table 4 .

For Chinese and English literature datasets in cloud computing, after completing preprocessing of relevant bibliographic information, the modeling files are input into a Java-based LDA program, and experiments are conducted following the hierarchical scientific knowledge structure discovery method described in Section 3.3. To obtain an optimal hierarchical knowledge structure, after multiple trials, we selected the threshold λ and minimum average similarity between topics S_0 from Table 5 as comparison parameters.

4.2.2 Experimental Results Analysis Based on the distribution of feature words under topics, potential knowledge topics are mapped layer by layer to obtain a three-layer knowledge structure in the cloud computing field. Partial results of the second-layer knowledge topics and their word probability distributions are shown in Tables 6 and 7 . Sorting feature words by probability distribution in descending order reveals that feature words under the same topic have strong relevance and similar semantic information, which greatly facilitates mapping of potential knowledge topics. We respectively draw knowledge structure hierarchy trees for Chinese and English corpora in the "cloud computing" field obtained by HSKSD and HLDA methods, as shown in Figures 6 [Figure 6: see original paper]-9 [Figure 9: see original paper].

Interpreting the knowledge structure generated by the HSKSD method, the first-layer knowledge topic for both Chinese and English is "Cloud Computing." The ten second-layer topics in Chinese include algorithm optimization, education domain, user services, data security, industrial innovation, storage processing, technology research, intelligent detection, information services, and platform architecture. The ten second-layer topics in English include Mobile network, Algorithm scheduling, Virtual machine, Image system, Cloud applications service, Datum security, Computing method, User resource service, Business management, and Data analysis. Consulting relevant materials, we find that these knowledge topics align with related concepts and technologies introduced in professional books on cloud computing and are basically consistent with previous research results in the "cloud computing" field, proving that our topic extraction results have certain scientific rationality. Analyzing the third layer of the knowledge structure reveals that, in addition to continuous attention to traditional technologies such as algorithms, storage, data security, and platform architecture, Chinese and English cloud computing research in recent years has involved various aspects including education, finance, transportation, and government affairs, with increasingly broad application scope. Cloud computing is essentially a collective term for a set of related technologies and services. Combined with cloud computing technologies and services, various emerging technologies such as IoT and image retrieval have attracted widespread scholarly attention, and various characteristic services such as recommendation services and intelligent monitoring services continue to emerge. Notably, the library and information field has also undergone some transformations, such as the transformation of libraries, publishing, and media industries.

Interpreting the knowledge structure generated by the HLDA method, the first-layer knowledge topic remains "Cloud Computing." Chinese second-layer knowledge topics include IoT, virtual machines, servers, databases, resource sharing, infrastructure, education domain, mobile communication, spatial information, data sources, and service outsourcing—eleven topics in total. English second-layer knowledge topics include system datum, mobile datum, computing performance, distributed database, network system, distribution strategy, and detection analysis—eight topics in total. These involve applications of cloud computing in education, communication, and other fields. Additionally, analyzing

the generated third-layer knowledge topics, "urban planning" and "genetic information" demonstrate cloud computing's contributions in smart cities and biomedicine. However, overall, the subordinate relationships among topics are relatively chaotic, such as "intellectual property" under "virtual machine" and "genetic information" under "mobile communication." Similar issues exist in English, such as "business innovation" and "privacy protection" under "computing performance."

Qualitative analysis of the above results is conducted from two aspects:

(1) Single-layer topic differentiation. Taking Chinese second-layer knowledge topics as an example, comparing knowledge topics generated by both methods reveals that the HSKSD method yields clearer knowledge topic classification, stronger inter-topic differentiation, and more accurate feature word representation, basically covering all aspects of the cloud computing field comprehensively and meticulously dividing it into ten research directions. In contrast, the HLDA method produces more scattered knowledge topics with overlapping areas, such as virtual machines and servers, databases and data sources, and fails to reflect relatively important aspects like data security, storage processing, and algorithms. Using English second-layer knowledge structure as another example, the HSKSD method similarly obtains ten knowledge topics with good structural relationships and clear topic differentiation, while the HLDA method yields unsatisfactory knowledge topics.

(2) Inter-layer topic inheritance. The main difference between hierarchical and traditional knowledge structure discovery lies in its hierarchical nature. By comparing topics across different layers, we observe whether good inheritance exists between layers and whether the hierarchy is clear. Using Chinese as an example, the third-layer knowledge topics corresponding to "algorithm optimization" in the second layer of the HSKSD method include network nodes, resource allocation, energy consumption analysis, target simulation, task scheduling, and load balancing—six aspects. These six topics have strong relevance and inheritance, representing finer-grained knowledge topics subordinate to "algorithm optimization." In contrast, the lower-level topics obtained by the HLDA method under "IoT" do not show obvious inheritance relationships, with knowledge points being relatively scattered. Using English as another example, the hierarchical topics obtained by the HLDA method similarly suffer from weak or even chaotic inheritance relationships.

4.3 Evaluation Metrics

The knowledge structure discovery method proposed in this paper is an unsupervised learning approach. Certain class labels or other benchmarks in the "cloud computing" field are not suitable as references, and traditional evaluation metrics such as accuracy, recall, and precision are not appropriate for validating our method's effectiveness. Therefore, inspired by relevant literature and beneficial discoveries during experiments, this paper comprehensively employs four eval-

uation metrics to assess the merits of HSKSD and HLDA methods: document utilization U (Utilization), document membership M (Membership), inter-topic independence I (Independence), and time complexity TC (Time Complexity). Symbols not mentioned earlier are explained as follows: (1) D represents the number of documents at layer i, where i $^{+}$; (2) D_T represents the number of documents under topic k at layer i; (3) word_{count} is the number of words in the post-processing vocabulary, i.e., the number of unique words.

The formulas for each metric are as follows:

$$U_i = \frac{\sum_{k=1}^{T_i} D\_T_{ik}}{D_i} \times 100\% \quad \text{Formula (1)}$$

$$M_i = \text{Avg}\left(\frac{\sum_{m=1}^{D\_T_{ik}} P(D_{im}, T_{ik})}{D\_T_{ik}}\right) \quad \text{Formula (2)}$$

$$I_i = 1 - \text{Avg}(S_i) = 1 - \frac{2\sum_{k=1}^{j-1}\sum_{r=k+1}^{j} \text{Sim}(T_{ik}, T_{ir})}{j(j-1)} \quad \text{Formula (3)}$$

$$TC_i = O_i(\text{niters} \times T_i \times \text{word\_count}_i) \quad \text{Formula (4)}$$

4.3.1 Document Utilization In experiments, whether all documents are assigned to the hierarchical structure is the first metric to measure. Document utilization U $\geq$ 100% is a reasonable range, where the formed hierarchical structure is comprehensive and complete, and generated topics can cover all documents. For Chinese, the second layer of the HSKSD method generates 10 different topics from all 6,115 documents. Summing documents under each topic yields 10,374 documents, greater than the total of 6,115 because one document can belong to different topics as long as its probability exceeds the intercept threshold λ. The same applies to English. In the HLDA method, each document searches for topics along a unique path, so the sum of literature equals the number of literature sets.

The document distribution under second-layer topics for Chinese and English is shown in Tables 8 and 9 . According to Formula (1) (with i = 2), document utilization U under different corpora can be obtained. Reviewing the threshold λ in Table 5 shows that threshold size is inversely proportional to document utilization; that is, larger thresholds make document utilization closer to 100%. This can be understood as higher thresholds reducing the number of selected articles under topics, thereby decreasing duplicate document assignments.

4.3.2 Document Membership All documents under a topic and their probability values P(D , T) of belonging to that topic constitute the "document-topic" matrix. According to Formula (2) (with i = 2), the average membership degree M of literature subsets under second-layer topics is calculated. Higher M values indicate that articles under the topic have higher membership and better clustering effects.

As shown in Figure 10 [Figure 10: see original paper], the M values of the HSKSD method are higher than those of the HLDA method under different corpora, indicating that our method forms topics with better clustering effects for documents.

4.3.3 Inter-topic Independence Comparing inter-topic independence essentially compares whether differences exist among feature words under topics. This paper uses feature words under each topic to calculate inter-topic independence I of the third-layer knowledge structure according to Formula (3) (with i = 3). Higher I values indicate greater inter-topic differences and lower coupling, with good clustering effects enhancing topic representation performance and model generalization ability.

During calculation, we found that similarity is high among lower-level topics under the same topic but low among lower-level topics under different topics, consistent with general inter-topic independence goals of high intra-cluster similarity and low inter-cluster similarity. Figure 11 [Figure 11: see original paper] shows inter-topic independence comparisons under different corpora for both methods.

The figure shows that the HLDA method generates stronger inter-topic independence. Analysis reveals that HLDA tends to select special, less frequent, and more distinctive words when choosing feature words, while the HSKSD method selects words with higher frequency in documents and broad representativeness as feature words, resulting in many identical feature words, which may cause lower inter-topic independence.

4.3.4 Time Complexity Algorithm time complexity is an important metric reflecting algorithm quality. Time complexity TC is commonly represented by symbol O; lower time complexity indicates higher efficiency. Under experimental environment configuration of MyEclipse 2015, JDK 1.8.0, 6GB memory, and Windows 7 operating system, we evaluated algorithm time complexity, as shown in Figure 12 [Figure 12: see original paper].

The complexity for each layer is: TC = O (niters × T × D × 1), where 1 is the average document length at each layer, i.e., 1 = word_{count} / D . Simplifying this formula yields Formula (4) (with i = 1, 2, 3). Figure 12 shows time complexity calculation results according to Formula (4), where iteration count niters is 1,000 for all cases.

The figure shows that under Chinese corpora, the HLDA method's time complexity is approximately 3 times that of HSKSD, while under English corpora it is nearly 4 times, with the HLDA method showing rapid complexity increase as layers deepen. Therefore, the HSKSD method is clearly superior to HLDA in terms of time complexity.

4.3.5 Comprehensive Comparison Comprehensive evaluation of the aforementioned comparison criteria shows that the proposed HSKSD method meets standards in document utilization comparable to HLDA, outperforms HLDA in document membership and time complexity, and is only inferior to HLDA in inter-topic independence. This comparison is from a quantitative perspective.

Additionally, reviewing the intuitive qualitative analysis of knowledge structures in Section 4.2.2, at the single-layer level, the HSKSD method discovers more comprehensive knowledge topics with less overlap and higher inter-topic differentiation compared to HLDA. At the multi-layer level, the HSKSD method shows stronger inter-layer inheritance, where lower-level knowledge represents finer-grained 刻画 of upper-level knowledge with good correlation or belonging.

Through the above analysis, comprehensive evaluation of the proposed HSKSD method and HLDA method under four quantitative metrics and two qualitative indicators yields the results shown in Table 10 .

References

[1] Du Bailan. On the construction principles of archivists' knowledge structure[J]. Beijing Archives, 1998(7): 28-29.

[2] Zhang Zhendong. The "wood" type discipline knowledge structure of intelligence personnel[J]. Journal of Intelligence Science, 1991(6): 475-478.

[3] Charvet FF, Cooper M C, Gardner J T. The intellectual structure of supply chain management: a bibliometric approach[J]. Journal of business logistics, 2008, 29(1): 47-73.

[4] Zhong Qiuyan, Qu Gang. Research on knowledge management discipline school division and development trends[J]. Information Science, 2011(1): 11-18.

[5] Song Ge. Comparative study of SNA and MSA in revealing knowledge structure[J]. Library and Information Service, 2009, 53(8): 106-109.

[6] Peng Xixian, Zhu Qinghua, Shen Chao. Author cooperation analysis in social computing field based on social network analysis[J]. Journal of Intelligence, 2013(3): 93-100.

[7] Zhou Z. Social network analysis of highly cited authors based on domestic mapping knowledge domains[J]. Journal of modern information, 2012, 32(8): 97-100.

[8] Rorissa A, Yuan X. Visualizing and mapping the intellectual structure of information retrieval[J]. Information processing & management, 2012, 48(1): 120-135.

[9] Li Wan, Sun Bindong. The knowledge structure and research hotspots of Western economic geography: a quantitative study based on CiteSpace maps[J]. Economic Geography, 2014, 34(4): 7-12.

[10] Ravikumar S, Agrahari A, Singh S N. Mapping the intellectual structure of scientometrics: a co-word analysis of the journal scientometrics (2005-2010)[J]. Scientometrics, 2015, 102(1): 929-955.

[11] Khasseh A A, Soheili F, Moghaddam H S, et al. Intellectual structure of knowledge in iMetrics[J]. Information processing & management, 2017, 53(3): 705-720.

[12] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003(3): 993-1022.

[13] Wang Yuefen, Fu Zhu, Chen Bikun. Exploring the research structure of domestic knowledge flow using LDA topic model: from the perspective of discipline classification topic extraction[J]. New Technology of Library and Information Service, 2016, 32(4): 8-19.

[14] Chang H C. The synergy of scientometric analysis and knowledge mapping with topic models: modeling the development trajectories of information security and cyber-security research[J]. Journal of information & knowledge management, 2016, 15(4): 77-84.

[15] Blei D M, Griffiths T L, Jordan M I, et al. Hierarchical topic models and the nested Chinese restaurant process[J]. Advances in neural information processing systems, 2004, 57(2): 18-22.

[16] Chen Jing, Xu Bo, Wang Tiantian, et al. Hierarchical organization of book internal topics based on hLDA[J]. Library and Information Service, 2016, 60(18): 140-148.

[17] Chen Liang, Zhang Jing, Zhang Haichao, et al. Application of hierarchical topic model in technology evolution analysis[J]. Library and Information Service, 2017, 61(5): 103-108.

[18] Zhang Yi, Shao Yudong, Zhang Jiawan. Visual analysis of topic evolution in multi-source media text[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(12): 2265-2272.

[19] Shanghai Linrun Information Technology Co., Ltd. HanLP[EB/OL]. [2017-08-10]. http://www.hanlp.linrunsoft.com/.

[20] The Stanford Natural Language Processing Group. Stanford CoreNLP[EB/OL]. [2017-07-10]. http://www.nlp.stanford.edu/software/corenlp.shtml.

[21] Wang Peng, Gao Cheng, Chen Xiaomei. Research on text clustering based on LDA model[J]. Information Science, 2015(1): 63-68.

[22] Heng Wei, Yu Jia, Li Lei, et al. Research on key factors of applying hLDA for multi-document topic modeling[J]. Journal of Chinese Information Processing, 2013, 27(6): 117-128.

[23] Blei D M, Griffiths T L, Jordan M I. The nested Chinese restaurant process and Bayesian inference of topic hierarchies[C]//International conference on neural information processing systems. Cambridge: MIT Press, 2012: 17-24.

[24] Erl T, Mahmood Z, Puttini R, et al. Cloud computing: concepts, technology & architecture[M]. Beijing: China Machine Press, 2014.

[25] Zhu Min. Review of current research status of cloud computing at home and abroad[J]. Computer Knowledge and Technology, 2015, 11(17): 52-53.

[26] Wang Jiandong, Liu Yang, Wang Jimin. Knowledge structure and evolution path analysis of core author groups in domestic cloud computing research field[J]. Journal of Peking University (Natural Science Edition), 2013, 49(5): 773-782.

[27] Li Huizong, Zhou Jiao, Wang Xiangqian, et al. User tag topic model fusing social relations[J]. Journal of Intelligence, 2017, 36(3): 165-172.

[28] Chen Min. Research on multimodal semantic knowledge base construction method[D]. Wuhan: Huazhong University of Science and Technology, 2014.

Author Contributions

Li Hui: Designed the research plan, proposed research ideas, and provided paper revision suggestions.
Tian Yadan: Responsible for experiment implementation, paper writing, and revision.

Abstract: [Purpose/significance] This paper proposes a new hierarchical discovery method for scientific knowledge structure, providing reference for optimizing the knowledge structure discovery process and improving knowledge organization forms. [Method/process] The method utilizes the LDA topic model to construct a hierarchical discovery approach for scientific knowledge structure. It automatically determines the number of hierarchical levels based on average similarity between topics, automatically filters thresholds in the "document-topic" probability matrix to intercept literature subsets for each topic, and finally employs tree diagrams to display scientific field knowledge structures, exploring correlations and inheritances among knowledge points, and compares with the hierarchical topic model HLDA method. [Result/conclusion] Through empirical research and comparison, the results show that the knowledge structure obtained by our method is better, with stronger knowledge topic representation

and higher operational efficiency, and demonstrates significant improvement over the HLDA method in single-layer topic differentiation and inter-layer topic inheritance.

Keywords: LDA; Cloud Computing; Hierarchical; Knowledge Structure

Note: Figure translations are in progress. See original paper for figures.